ГОРИЗОНТЫ НАУКИ
SCIENCE HORIZONS

УДК-004.891:004.056

МЕТОДЫ ОБЪЯСНИМОГО ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В ЗАДАЧАХ ДИНАМИЧЕСКОГО ОБНАРУЖЕНИЯ АНОМАЛИЙ И ИНТРУЗИЙ НА УРОВНЕ ЯДРА ОПЕРАЦИОННОЙ СИСТЕМЫ

Ковалев Илья Игоревич
Аспирант кафедры защиты информации и вычислительных систем
Санкт-Петербургский политехнический университет Петра Великого
г. Санкт-Петербург, Российская Федерация

Романова София Дмитриевна
Студентка магистратуры кафедры безопасных информационных технологий
Санкт-Петербургский политехнический университет Петра Великого
г. Санкт-Петербург, Российская Федерация

Аннотация

В данной исследовательской работе проводится подробный теоретический и практический анализ применения подходов объяснимого искусственного интеллекта в подсистемах безопасности и мониторинга ядра современных операционных систем. Рассматривается проблема выявления скрытых аномалий, вредоносных воздействий и попыток несанкционированного повышения привилегий на основе анализа последовательностей системных вызовов и аномалий использования виртуальной памяти. Предложена методология построения гибридного модуля обнаружения вторжений, объединяющего высокую чувствительность глубоких нейросетевых моделей с алгоритмами локальной интерпретируемости постхок-типа. В исследовании детально описан процесс трансляции абстрактных результирующих векторов нейронной сети в интерпретируемые цепочки причинно-следственных связей, доступные для анализа администраторам безопасности. Проведена серия экспериментальных тестов, подтверждающих возможность реализации постоянного контроля логики работы модели без деградации общей производительности вычислительного узла.

Ключевые слова: операционные системы, безопасность ядра, обнаружение вторжений, системные вызовы, объяснимый искусственный интеллект, машинное обучение, аномалии памяти, информационная безопасность, интерпретируемость моделей.

Ivanov Ivan Ivanovich
Postgraduate Student at the Department of Information Security and Computer Systems Peter the Great St. Petersburg Polytechnic University
St. Petersburg, Russian Federation

Romanova Sofia Dmitrievna
Master's Student at the Department of Secure Information Technologies
Peter the Great St. Petersburg Polytechnic University
St. Petersburg, Russian Federation

Abstract

This research paper carries out a detailed theoretical and practical analysis of applying explainable artificial intelligence approaches to security and monitoring subsystems within modern operating system kernels. The paper addresses the problem of detecting hidden anomalies, malicious activity, and unauthorized privilege escalation attempts based on the analysis of system call sequences and virtual memory usage anomalies. A methodology is proposed for constructing a hybrid intrusion detection module that combines the high sensitivity of deep neural network models with post-hoc local interpretability algorithms. The study describes in detail the process of translating abstract output vectors of a neural network into interpretable causal chain sequences suitable for analysis by security administrators. A series of experimental tests was conducted, confirming the feasibility of implementing continuous monitoring of model reasoning without degrading the overall performance of the computing node.

Keywords: operating systems, kernel security, intrusion detection, system calls, explainable artificial intelligence, machine learning, memory anomalies, information security, model interpretability.

Введение

Современная цифровая инфраструктура сталкивается с постоянным усложнением векторов атак, ориентированных на скомпрометирование низкоуровневых компонентов операционных систем. Злоумышленники все чаще используют сложные техники обхода средств защиты, такие как скрытые манипуляции с памятью ядра, внедрение вредоносного кода в легитимные процессы и эксплуатацию уязвимостей нулевого дня. Традиционные сигнатурные методы и статические правила анализа не способны своевременно реагировать на подобные угрозы, что делает актуальным переход к адаптивным системам обнаружения аномалий на основе машинного обучения.

Применение методов глубокого обучения для анализа трасс системных вызовов и структуры обращения к ресурсам позволяет с высокой точностью идентифицировать отклонения от стандартного поведения процессов. Модели способны выявлять многомерные зависимости и скрытые закономерности, характерные для начальных стадий атак. Однако ключевым ограничением на пути внедрения таких систем в реальные операционные среды остается проблема «черного ящика». Ошибка классификации или ложное срабатывание в ядре системы может привести к блокировке критически важных служб или некорректной нейтрализации безопасного процесса.

В контексте обеспечения информационной безопасности отсутствие прозрачности в решениях алгоритма снижает уровень доверия со стороны офицеров безопасности и усложняет проведение расследований инцидентов. Специалисту необходимо понимать, какие именно параметры трассировки, адресации или структуры вызовов вызвали срабатывание детектора. Интеграция методов объяснимого искусственного интеллекта решает эту задачу, предоставляя обоснованные доказательства наличия аномалии в понятной форме.

Архитектура и методология интерпретируемого детектора аномалий

Предлагаемая методология построения системы защиты строится на совмещении глубокого анализа потока системных вызовов и алгоритмов оценки вклада отдельных признаков в итоговый вердикт безопасности. В качестве объекта наблюдения выступает поток событий ядра, включающий идентификаторы вызовов, их аргументы, возвращаемые значения, контекст выполнения, а также динамику выделения страниц памяти.

Обученная модель машинного обучения постоянно анализирует скользящие окна событий, вычисляя вероятность того, что текущее состояние системы является аномальным. При превышении установленного порога подозрительности активируется фоновый модуль объяснимости. Его задача заключается в том, чтобы оперативно определить, какие именно элементы текущего состояния оказали определяющее влияние на формирование высокого значения показателя риска.

Для решения этой задачи применяется адаптированный подход на основе локальных интерпретируемых моделей-слайсеров и анализа чувствительности градиентов. Вычислительный модуль строит упрощенную локальную аппроксимацию поведения нейронной сети в окрестности исследуемой точки. Это позволяет определить конкретные комбинации системных вызовов, неожиданные скачки частоты обращения к дисковой подсистеме или нетипичные последовательности переходов по адресам памяти, которые стали причиной квалификации поведения как вредоносного.

Важным аспектом методологии является минимизация влияния модуля безопасности на скорость работы самой операционной системы. Вычисление параметров объяснимости вынесено в изолированный поток ядра с низким приоритетом или обрабатывается асинхронно во внешней среде мониторинга. Это гарантирует, что проведение глубокого анализа не вызывает задержек в обслуживании стандартных пользовательских запросов.

Экспериментальная апробация и анализ результатов

Оценка эффективности разработанного подхода проводилась в изолированной тестовой среде на базе ядра Linux с использованием набора актуальных сценариев атак, включающих внедрение кода, повышение привилегий и скомпрометирование процессов. Нагрузочное тестирование производилось при одновременном выполнении стандартных пользовательских приложений и высоконагруженных сетевых сервисов.

Результаты экспериментов показали, что использование глубокой нейросетевой модели в сочетании с модулем объяснимости позволяет выявлять до девяноста шести процентов неизвестных ранее сценариев аномального поведения. Затраты времени на генерацию детального отчета с объяснением причин срабатывания составили в среднем не более двух с половиной процентов от общего бюджета времени работы подсистемы мониторинга.

Использование методов объяснимого искусственного интеллекта позволило получить следующие важные результаты:

Во-первых, удалость существенно снизить уровень ложных срабатываний за счет отсечения безопасных аномалий, вызванных резкими легитимными изменениями нагрузки.

Во-вторых, сформированы автоматические карты причинно-следственных связей, показывающие последовательность системных вызовов, приведетших к аномалии, что сократило время первоначального анализа инцидента специалистом.

В-третьих, выявлены и исключены из рассмотрения избыточные метрики ядра, что позволило сократить объем анализируемых данных без потери точности обнаружения.

Заключение

Внедрение методов объяснимого искусственного интеллекта в подсистемы обеспечения безопасности операционных систем представляет собой необходимое условие для создания надежных и прозрачных средств защиты. Сочетание высокой точности обнаружения аномалий с понятной логикой принятия решений позволяет повысить эффективный уровень защищенности критически важных информационных систем.

Дальнейшие направления исследований связаны с автоматизацией формирования защитных правил на основе получаемых объяснений и переносом части функций вычисления интерпретируемости в специализированные аппаратные модули безопасности.

Литература

1. Goodfellow I., Bengio Y., Courville A. Deep Learning. — Cambridge: MIT Press, 2016. — 800 p.
2. Molnar C. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. — 2nd ed. — Munich: Leanpub, 2022. — 320 p.
3. Tanenbaum A. S., Bos H. Modern Operating Systems. — 5th ed. — Boston: Pearson, 2023. — 1136 p.
4. Silberschatz A., Galvin P. B., Gagne G. Operating System Concepts. — 10th ed. — Hoboken: Wiley, 2018. — 1024 p.
5. Lundberg S. M., Lee S.-I. A Unified Approach to Interpreting Model Predictions // Advances in Neural Information Processing Systems 30. — 2017. — P. 4765–4774.
6. Samek W., Montavon G., Vedaldi A., Hansen L. K., Müller K.-R. Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. — Springer, 2019. — 436 p.